Lab 12: confounders, mediators, suppressors, and path diagrams

Statistics for Social Scientists II

Bur, GJM

2024-11-25

Confounding

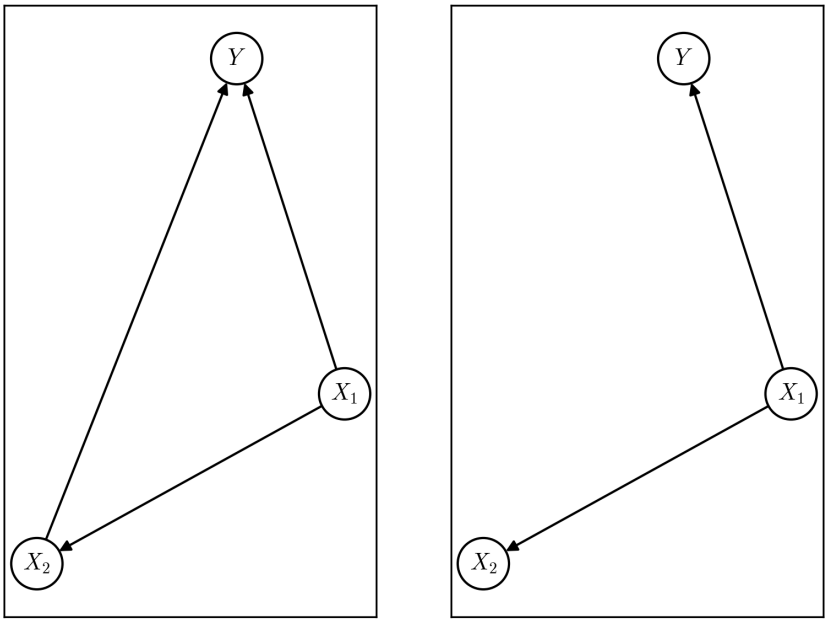

Figure 1: X_2 mediates the effect of X_1 on Y; or, X_1 confounds X_2 and Y

1 Confounding

Confounding is a common problem. It basically means, as in the diagram shown below, that X_1 causes both X_2 and Y, which have no direct relation; so, there appears to be a

relationship between X_2 and Y but there is none. It's easy to think of examples. For example, one might theorize that education largely has a signalling effect (meaning, people largely get education because they are capable; education does not actually add to their ability), and if we could observe ability independent of education, a regression of wages on ability and education would show no effect of education.

By the way, confounding does not require that the relations be causal. It could be the case that we *know* that, if anything, X_2 causes X_1. For example, it's possible that education partially causes ability. However, it is also possible that education only matters through the ability channel and other things also cause ability. In this case, including ability in the model causes education to become insignificant (and if you were completely certain that education caused ability, you would want to consider a mediation analysis). In practice, it is often difficult to say which of these situations actually obtains because the theory may not provide an inarguable answer.

Let's look at the GSS and let's look at income. First, here is an example of how you might clean data realistically on the GSS.

```
cd ~/desktop/code/soc361fa24
use ./data/gss2018, clear
keep race educ rincom16 sex class hispanic age maeduc paeduc padeg madeg

* Make an income variable with automatic midpoint imputation
gen inc_midpt = .

    * As a first step, let's replace numeric as missing and make a clearer
    * value label for rincom91==1.

    replace inc_midpt = . if rincom16 > 26 // 27 and up is missing
    replace inc_midpt = 170000 if rincom16 == 26

    lab def rincom16 1 "$0-1000", modify

    levelsof rincom16, local(levs)
        * This stores all possible values of the variable in a local
        * called "levs"
    local inclab : value label rincom16
        * We also put the value label itself into a local, using Stata's
        * extended macro functions, which let us use the syntax above (for
        * more on this, see p. 4 of this and on:
            * https://www.stata.com/manuals13/pmacro.pdf)

    forvalues l = 2/25 {
        di `l'
        local strlab : label `inclab' `l'
            * What's happening here? We access the string label associated
                * with numeric value l, for l in the set of all possible
                * levels of income
        local strlab = subinstr("`strlab'", " ", "", .)
```

```
            * We have different numbers of spaces in some of the numbers,
            * so it's easier to just make it one long string
        local strlab = subinstr("`strlab'", "$", "", .)
            * We also have inconsistent use of dollar signs so let's just
            * delete them
        di "`strlab'"
        local n1 = strrpos("`strlab'", "to") - 1
        * Now, I want to turn that string into two different numbers
        * but the problem is that the numbers are of variable length
        * so I just say "tell me at what position in the string the
        * word 'to' first occurs" and store that as n21 I subtract one
        * for reasons mentioned below
        local lb = substr("`strlab'", 1, `n1')
        * Now, I get the lower bound of the interval by taking the sub-
        * string from the second position of the string label for a
        * length of n1. Now we see why I took away one above in defining
        * n1: I didn't want to include the word itself, and I was
        * starting the string after the first character, a dollar sign
        local ub = substr("`strlab'", `n1'+3, strlen("`strlab'"))
        * Now I get the ub. I add three to n2 to start the second string
        * after the hyphen and then go until the end of the string label
        * using the length of the string

        di "`lb'"
        di "`ub'"
        replace inc_midpt = round((real("`lb'") +  real("`ub'"))/2, 1)  ///
            if rincom16 == `l'
        * Finally, I replace income with the average of my two values
        }
lab var inc_midpt "Income (real, midpoint imputation)"
tabstat inc_midpt, by(rincom16)

* Make a race variable with hispanic included
* assuming that Hispanic is a separate category
* from white and black.
gen wbho = race
replace wbho = 4 if hispanic >1 & !missing(hispanic)
label copy RACE wbho
lab val wbho wbho
label define wbho 4 "hispanic", modify
reg inc i.wbho i.sex c.educ i.class
```

Let's now enter in some variables into our model.

```
reg inc_midpt i.wbho
```

Let's enter in a few controls.

```
reg inc i.wbho educ i.class
```

How do we interpret this? Surely class and education cannot "cause" race. The way to interpret this is that race is a proxy for the effects of class and education. One way to think about it is that class and education affect income, and being hispanic affects income largely through this channel, but other things determine class besides race.

By the way, note that in this case, we have an interesting pattern of effects where entering education or class on its own into the model is not enough to cause the hispanic dummy to become insignificant, but both together are; this is common enough and again requires the researcher to use some judgment.

Let's look at another example. Suppose we have a model that suggests that for respondents who were adults in 2018, father's education was the actual means by which status was transmitted to children, with mother's education not mattering as much directly but being selected for by the father—perhaps a plausible theory in light of feminist claims about patriarchal societies (this theory itself might be patriarchal, in its own way; that debate is a tale for another time).

Check this theory using the GSS.

2 Mediators, moderators, suppressors

If we have the structure of relationships shown above, we could either consider the relationship between X_2 and Y to be totally spurious or to be a case of **mediation**.

In the case of a totally spurious relationship, we might think that there is no qualitative sense in which X_2 is involved in the chain of causality. For example, firework sales go up when people wear bathing suits. In a strict sense, there is no causal relationship here at all; both are caused by the summer time in the United States. There is not much point in closely investigating the relationship of fireworks and bathing suits once we know this.

However, there might be a situation where we think that the causal relationship between X_1 and Y *is* mediated by X_2. For example, in the case of education and ability, if ability is ultimately responsible for education but this is in some sense a meaningful expression (perhaps, even though most education is just ceremonial, almost all talented people do it anyways), this can be of interest to analyis. In this setting, we call education a *mediator*.

Don't confuse these mediators with **moderators**, variables that essentially have interactions with each other.

Suppressors are like mediators but when the mediator has an opposite effect on the outcome.

Let's see how we can compute the **total**, **direct**, and **indirect effects** of mediators/suppressors.

3 Path analysis

If we write $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ and then let $X_2 = \beta_{210} + \beta_{211} X_1 + \epsilon$—the logic of the notation is that the first two numbers in the subscript tell us what is regressed on what, then the third is the number from the standard way of distinguishing regression coefficients—then we can simply swap in:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2(\beta_{210} + \beta_{211} X_1 + \epsilon) + \epsilon$$

Then, identifying the effect of β_1 is even simpler. I'm going to use calculus notation since I greatly prefer it, and we have now finally introduced it. Note that $f'(x) = \frac{dy}{dx}$ and both mean "the instantaneous rate of change of y with respect to x"; this is called the derivative, and it is also the slope of the tangent line at any particular x_0, assuming the derivative includes an x-term—as did, for example, our quadratic function, $y = ax^2 + bx + c$, whose derivative is the most famous of them all, $2ax + b$. If our function is *linear*, the derivative is constant across all values of x but it is still hand notation to mean "the rate of change of y with respect to x for small changes", whether or not those are constant across x. The notation $\frac{\partial y}{\partial x}$ means a *partial* derivative and simply means that there could be other inputs $z, w, u, v...$ that we are holding constant.

$$\frac{\partial Y}{\partial X_1} = \beta_1 + \beta_2 \beta_{211}$$

We call β_1 the direct effect, $\beta_2\beta_{211}$ the indirect effect, and their sum $\beta_1 + \beta_2\beta_{211}$ the total effect.

Calculate the direct, indirect, and total effect of father's education on respondent's education assuming that mother's education is the mediator. Verify that this is the same thing as the regression slope given by the bivariate regression of income on father's education.

Something like this should work:

```
drop if missing(maeduc, paeduc, inc)
reg inc paeduc maeduc
local te = round(_b[paeduc], 0.01)
local B_incm = _b[maeduc]
di "Direct effect of paeduc is `te'"
reg maeduc paeduc
local B_mp = _b[paeduc]
local ide = `b_mp'*`macoeff'
di "Indirect effect of paeduc is `ide'"
local de = `te' + `ide'
di "Direct effect is `de'"
reg inc paed, noh
```

3.1 Perspectives on multiple regression

Path analysis points out a fact that is pretty important, that the multiple regression slope β_2 from the regression $Y = \beta_1 X_1 + \beta_2 X_2$ (assume centered variables for simplicity) is not just the regression slope from $Y = \beta_2 X_2$. Instead, if we want to think about how to do

multiple regression sequentially, we would first regress Y on X_1 and obtain the residuals. These represent variation in Y unexplained by X_1; call these $\hat{\epsilon}_1$. Then, regress X_2 on X_1. Now, the residuals represent variation in X_2 that is not explicable by means of X_1; call these $\hat{\epsilon}_2$. The regression of $\hat{\epsilon}_1$ on $\hat{\epsilon}_2$ happens to give the same slope as that on X_2 from the full model. *This fact is sometimes called the Frisch-Waugh-Lovell theorem.* **Try this out with a regression of income on age and education; find the slope on age from the full model by this method**.

Something like this should work.

```
reg inc educ
predict e1, residuals
reg age educ
predict e2, residuals
reg e1 e2
reg inc educ age, nohe
```

3.1.1 The subject space proof of the Frisch-Waugh-Lovell Theorem

There is also nice geometric proof in subject space that I'll show below (forgive my bad handwriting; I did not have time to code this). I'm not sure why Gordon mentions it in this chapter; Wright's work has some interesting perspectives on what multiple regression really *is*, but he doesn't really give an intuition for anything like the FWL theorem, and neither does Gordon. This proof instead is inspired by (but not fully spelled out) in the work of Wickens (1995).[1]

In what follows, I used the more technically-correct notation of U for errors and $\hat{u}$ for residuals. The idea is as follows: the sample regression of Y on X_1 gives a residual vector $\hat{u}_1$ that has the same S_2 coordinate (what we would normally call the y-axis, avoided here to evade confusion) as Y, and the same S_1 coordinate (0, on what we would normally call the x-axis). The sample regression of X_2 on X_1 gives a residual vector $\hat{u}_2$ with the same S_2 coordinate as X_2 and the same S_1 coordinate (again, 0). So regressing $\hat{u}_1$ on $\hat{u}_2$ requires us to shrink the longer vector, $\hat{u}_1$ by the same amount that we would need to shrink X_2 by in the multiple regression to get it to the height of Y's projection into the X_1X_2 plane. So, by a similar triangles argument, the slopes must be the same.

4 Appendix

Wright's original paper shows some interesting extensions.[2] Let all of our variables be standardized, let z_0 be the outcome and $z_2, z_3, ...z_m$ be the predictors. Let our model be …

[1] *The Geometry of Multivariate Statistics*, a book I've mentioned often. It is the best book on statistics ever written for non-statisticians, I think.

[2] When I took this course, I noticed that it was an odd coincidence that probably our department's most famous (at the time) member was Erik Olin Wright, its previous probably most famous member William H. Sewell, Sr. (after whom our building is named), and that Sewall Wright, also rather famous, taught here. I asked Erik Wright about this; it was apparently a common question, but there is no apparent connection. There is also an unrelated and quite prominent Olin Park here in town, just south of downtown on John Nolen, just to add to the mystery.

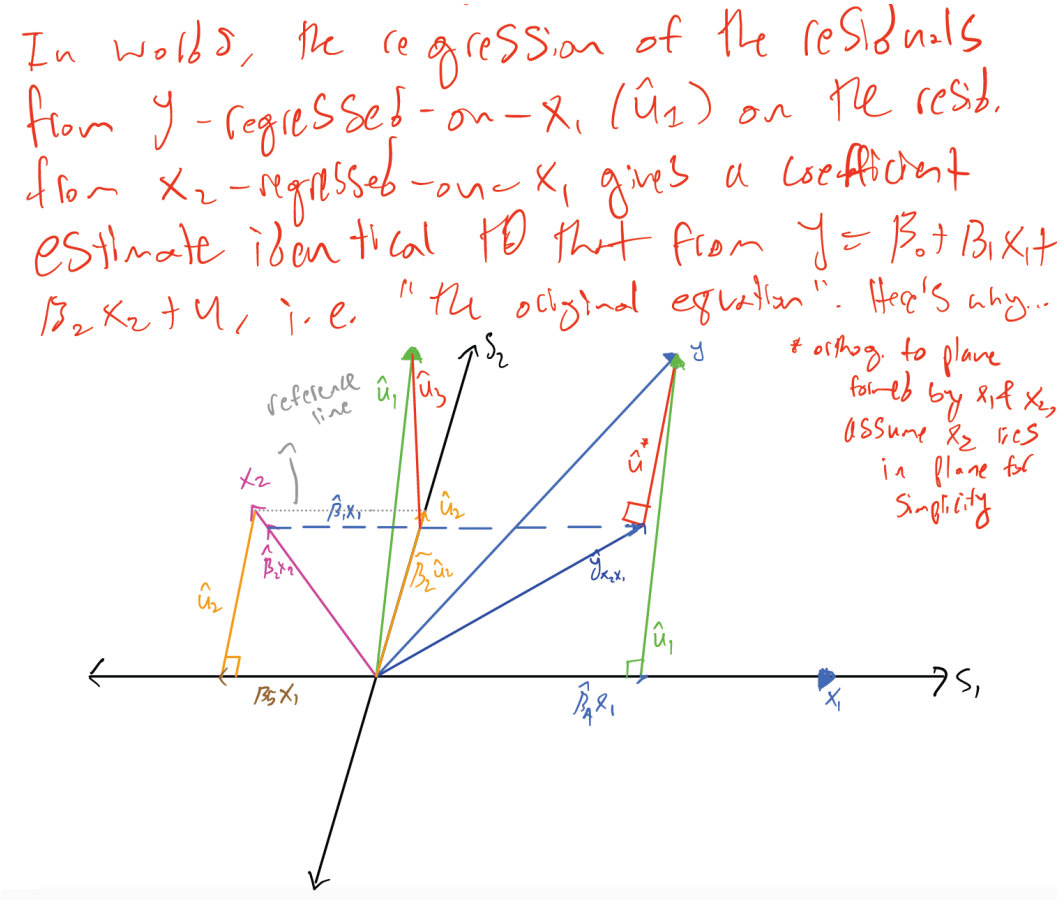

Figure 2: FWL_illustrated

$$z_0 = \sum_{k=1}^{m} P_k z_k + \epsilon$$

… where the P_k are the *p*ath coefficients, which are the standardized regression slopes that would come from the following model: $Y = \beta_0 + \sum_{k=1}^{m} \beta_k X_k + \epsilon$, where z_0 represents the standardized version of y and all other connections are obvious.[3]

Then, the correlation between any variable z_q and the outcome is…

$$\begin{aligned}
\rho_{0k} &= \frac{1}{N} \sum_{i=1}^{N} z_{0i} z_{qi} \\
&= \frac{1}{N} \sum_{i=1}^{N} \left\{ \sum_{k=1}^{m} P_k z_{ki} \right\} z_{qi} \\
&= \sum_{i=1}^{N} \sum_{k=1}^{m} P_k \frac{z_{ki} z_{qi}}{N} \\
&= \sum_{k=1}^{m} P_k \rho_{kq}
\end{aligned}$$

[3] I am cleaning up Wright's (1934) notation here; he uses them in a way that might be a little confusing (e.g. using V_k to represent the original variables and X to represent standardized versions.